

Faculty of Engineering and Information Technology  
University of Technology, Sydney

**Outlier Detection in Large  
High-Dimensional Data and Its  
Application in Stock Market  
Surveillance**

A thesis submitted in partial fulfillment of  
the requirements for the degree of  
**Doctor of Philosophy**

by

Chao Luo

February 2011

## CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

  
\_\_\_\_\_

# Acknowledgments

I would like to express my deep and sincere gratitude to my principal supervisor, Professor Chengqi Zhang, for helping me to get scholarship and offering me the chance to this study. Without his support, this thesis would be impossible. He has given me valuable guidance on my research during my doctoral study. He has provided opportunities for me to join industry projects which made me accumulate professional experience and knowledge.

I am deeply indebted to my co-supervisor Dr. Yanchang Zhao. The successful completion of this dissertation is due in great part to his support and guidance. I appreciate all his contributions of time, ideas and advices to help me complete this research. During the past three years, he has discussed with me regularly. He helped me develop ideas, design experiments, write academic papers and so on. I learned not only the skills of academic research, but also the right attitude on research from him. He taught me how to do interesting and meaningful research.

I am grateful for Prof. Longbing Cao for his suggestions and guidance on my research. He has provided me the good research environment and resources during my PhD study.

I am also grateful to my family for their continuous support in life. My parents, my brother and my sisters encouraged me and supported me whenever I met difficulty. Their love for me gave me strength to face difficulties over the years.

My additional thanks goes to the following schoolfellows for their suggestions and help during my study: Dr. Yuming Ou, Dr Jiarui Ni, Mr Zhigang

Zheng, etc., all members in the Data Sciences and Knowledge Discovery Research Lab, as well as all relevant staff and students in the Faculty and the University Graduate School. During my doctoral study, they gave me many help on study.

Finally, I appreciate the support from the CMCRC Scholarship for my research, which contributes to the delivery of this thesis.

# Contents

Certificate . . . . .	i
Acknowledgment . . . . .	ii
List of Figures . . . . .	vii
List of Tables . . . . .	ix
Abstract . . . . .	x
 <b>Chapter 1 Introduction . . . . .</b>	 <b>1</b>
1.1 Stock Market Surveillance . . . . .	1
1.2 Outlier Detection on Multiple Stock Data . . . . .	2
1.3 Agent-based Subspace Clustering . . . . .	3
1.4 Outlier Detection by Subspace Clustering . . . . .	4
1.5 Contributions . . . . .	5
1.6 Organization of The Thesis . . . . .	5
 <b>Chapter 2 Background and Literature Review . . . . .</b>	 <b>7</b>
2.1 Stock Market Surveillance . . . . .	7
2.1.1 Stock Market . . . . .	7
2.1.2 Stock Surveillance . . . . .	8
2.1.3 Market Surveillance Process . . . . .	10
2.2 Technologies for Stock Market Surveillance . . . . .	13
2.2.1 Rule Approaches . . . . .	13
2.2.2 Basic Statistic Methods . . . . .	14
2.2.3 Outlier Identification . . . . .	14
2.2.4 Other Techniques . . . . .	17

2.3	Clustering High-Dimensional Data . . . . .	24
2.4	Outlier Detection in Large High-Dimensional Data . . . . .	27
2.5	Summary . . . . .	30
<b>Chapter 3</b>	<b>Outlier Detection on Multiple Stock Data . . . . .</b>	<b>31</b>
3.1	Introduction . . . . .	31
3.2	Outlier Detection on Multiple Stock Data . . . . .	31
3.2.1	V-BOMM Model . . . . .	32
3.2.2	P-BOMM Model . . . . .	33
3.2.3	Description of V-BOMM and P-BOMM . . . . .	34
3.3	A Case Study in Stock Market Surveillance . . . . .	36
3.3.1	Data Sets . . . . .	36
3.3.2	Experiments Setup . . . . .	37
3.3.3	Evaluation Metrics . . . . .	38
3.3.4	Experimental Results . . . . .	39
3.4	Conclusion . . . . .	45
<b>Chapter 4</b>	<b>Agent-based Subspace Clustering . . . . .</b>	<b>46</b>
4.1	Introduction . . . . .	46
4.2	Agent-based Subspace Clustering . . . . .	47
4.2.1	Problem Statement . . . . .	47
4.2.2	Agent-based Subspace Clustering . . . . .	48
4.2.3	Algorithm . . . . .	52
4.3	Experiments . . . . .	54
4.3.1	Data and Evaluation Criteria . . . . .	54
4.3.2	Experimental Steps . . . . .	56
4.3.3	Experimental Results . . . . .	56
4.4	A Case Study in Stock Market Surveillance . . . . .	62
4.5	Conclusion . . . . .	65
<b>Chapter 5</b>	<b>Outlier Detection by Subspaces Clustering . . . . .</b>	<b>67</b>
5.1	Introduction . . . . .	67

5.2	Reference-based Outlier Detection by Subspace Clustering . .	67
5.2.1	General Idea . . . . .	67
5.2.2	Agent-based Subspace Clustering . . . . .	68
5.2.3	Outlier Detection in Subspaces . . . . .	69
5.2.4	Algorithm . . . . .	71
5.3	Experiments . . . . .	71
5.3.1	Data Sets . . . . .	71
5.3.2	Experimental Results . . . . .	73
5.4	A Case Study in Stock Market Surveillance . . . . .	81
5.5	Conclusion . . . . .	85
<b>Chapter 6 Conclusions and Future Work . . . . .</b>		<b>86</b>
6.1	Conclusions . . . . .	86
6.2	Future Work . . . . .	88
<b>Appendix A List of Publications . . . . .</b>		<b>93</b>
<b>Bibliography . . . . .</b>		<b>95</b>

# List of Figures

3.1	Design of OMM . . . . .	33
3.2	VOMM on Daily Price Return . . . . .	40
3.3	VOMM on Daily Price Range . . . . .	40
3.4	VOMM on Daily Trade Amount . . . . .	41
3.5	Probability-Based OMM . . . . .	41
3.6	Comparison on Accuracy of Different Models . . . . .	43
3.7	Comparison on Precision of Different Models . . . . .	43
3.8	Comparison on Specificity of Different Models . . . . .	44
3.9	Comparison on Recall of Different Models . . . . .	44
4.1	An Example of Bin . . . . .	49
4.2	An Example of CLIQUE Clustering . . . . .	50
4.3	An Example of Agent-based Clustering . . . . .	52
4.4	F1 Measure on Breast Data . . . . .	57
4.5	Entropy on Breast Data . . . . .	57
4.6	F1 Measure on Diabetes Data . . . . .	57
4.7	Entropy on Diabetes Data . . . . .	58
4.8	F1 Measure on Glass Data . . . . .	58
4.9	Entropy on Glass Data . . . . .	58
4.10	F1 Measure on Pendigits Data . . . . .	59
4.11	Entropy on Pendigits Data . . . . .	59
4.12	F1 Measure on Liver Data . . . . .	59
4.13	Entropy on Liver Data . . . . .	60



4.14	F1 Measure on Shape Data . . . . .	60
4.15	Entropy on Shape Data . . . . .	60
4.16	Scalability with Dimensionality . . . . .	61
4.17	Scalability with Data Size . . . . .	62
4.18	F1 Measure on Stock Data . . . . .	63
4.19	Entropy on Stock Data . . . . .	64
4.20	Running Time with $\xi$ . . . . .	65
4.21	Running Time with $\tau$ . . . . .	65
5.1	Precision on Segment Data . . . . .	75
5.2	Accuracy on Segment Data . . . . .	76
5.3	Specificity on Segment Data . . . . .	76
5.4	Recall on Segment Data . . . . .	77
5.5	Precision on Pendigits Data . . . . .	77
5.6	Accuracy on Pendigits Data . . . . .	78
5.7	Specificity on Pendigits Data . . . . .	78
5.8	Recall on Pendigits Data . . . . .	79
5.9	ROC Curve on Segment Data . . . . .	80
5.10	ROC Curve on Pendigits Data . . . . .	80
5.11	Scalability with Dimensionality . . . . .	81
5.12	Precision on Stock Data . . . . .	82
5.13	Accuracy on Stock Data . . . . .	83
5.14	Specificity on Stock Data . . . . .	83
5.15	Recall on Stock Data . . . . .	84
5.16	ROC Curve on Stock Data . . . . .	84

# List of Tables

2.1	Market Manipulation . . . . .	11
3.1	Examples of Original Alerts . . . . .	37
3.2	Comparison on the Number of Correctly Detected Outliers . .	39
4.1	A simple Example of Data Points . . . . .	47
4.2	Data Sets After Discretization . . . . .	50
4.3	Public Data Sets from UCI Repository . . . . .	54
4.4	Algorithms for Clustering . . . . .	55
5.1	Public Data Sets . . . . .	73
5.2	Algorithms for Outlier Detection . . . . .	74

# Abstract

Outlier detection techniques play an important role in stock market surveillance that involves analysis of large volume of high-dimensional trading data. However, outlier detection in large high-dimensional data is very challenging and is not well addressed by existing techniques. Firstly, it is difficult to select useful and relevant features from high-dimensional data. Secondly, large high-dimensional data need more efficient algorithms.

To attack the above issues brought by large high-dimensional data, this thesis presents two outlier detection models and one subspace clustering model.

Firstly, an outlier mining model is proposed to detect the outliers from multiple complex stock market data. In order to improve the efficiency of outlier detection, a financial model is used to select the features to construct multiple datasets. This model is able to improve the precision of outlier mining on individual measurements. The experiments on real-world stock market data show that the proposed model is effective and outperforms traditional technologies.

Secondly, in order to find relevant features automatically, an agent-based algorithm is proposed to discover subspace clusters in high dimensional data. Each data object is represented by an agent, and the agents move from one local environment to another to find optimal clusters in subspaces. Heuristic rules and objective functions are defined to guide the movements of agents, so that similar agents(data objects) go to one group. The experimental results show that our proposed agent-based subspace clustering algorithm performs

better than existing subspace clustering methods on both F1 measure and Entropy. The running time of our algorithm is scalable with the size and dimensionality of data. Furthermore, an application of our technique to stock market surveillance demonstrates its effectiveness in real world applications.

Finally, we propose a reference-based outlier detection model by agent-based subspace clustering. At first, agent-based subspace clustering is utilized to generate clusters in subspaces. After that, the centers of clusters, together with the corresponding subspaces, are used as references, and a reference-based model is employed to find outliers in relevant subspaces. The experimental results on real-world datasets prove that the proposed model is able to effectively and efficiently identify outliers in subspaces.

In summary, this thesis research on outlier detection techniques on high-dimensional data and its application in stock market surveillance. The proposed models are novel and effective. They have shown their potentials in real business.